



UNCLASSIFIED

Information Science and Technology Seminar Speaker Series and Data Science at Scale Summer School Speaker Series



Yanif Ahmad
The Johns Hopkins University

Towards Data On Tap: Data Views for Continuous Data Workflow and Exploration Systems

Wednesday, August 6, 2014

3:00 - 4:00 PM

TA-3, Bldg. 1690, Room 102 (CNLS Conference Room)

Abstract: Many science and big data domains are increasingly developing and deploying custom data processing and analysis pipelines that act as long-lived foundational infrastructure. These data pipelines are enabling technologies that drive community formation; individual researchers benefit from the economies of scale in shared software and hardware, and in analysis capabilities. Today, data pipelines are developed organically with a myriad of tools, from simulation codes, to scripting languages and relational databases. We propose structured pipeline design through database-style views. Data views are a key abstraction mechanism in databases that separate the definition of a data product from the computation of the data product, providing data encapsulation and independence at each workflow stage. This talk presents our efforts at realizing data pipeline development that uses views as a building block. Our DBToaster project enables highly-efficient incremental maintenance of data views as new data arrives at long-running pipelines. DBToaster uses novel SQL query compilation techniques to create lightweight, specialized view maintenance engines for streaming data transformation and aggregation, with 2-4 orders of magnitude speedup over DBMS and data stream systems. Our ongoing Mosaic project realizes a scalable implementation of DBToaster by sharding data views across main-memory available on commodity compute clusters. Finally, our BigDig project leverages views for joint simulation and analysis aimed at exploring high-dimensional simulation datasets, with applications to data-driven control of molecular dynamics codes. Time permitting, I will briefly present a teaser on BigDig's proposed combinatorial views for automatic feature exploration in simulation datasets.

Biography: Ahmad studies and designs novel abstractions for large-scale data management. He is affiliated with the Data Management Systems Lab; Computer Systems Research Group and the Institute for Data-Intensive Engineering and Science. His research spans foundations and applications, with the K3 project realizing programming abstractions for declarative, democratized construction of distributed data systems, and the Molecular Dynamics Database pursuing data-intensive computing architectures and analytics for large biological datasets. Ahmad received his Ph.D. from Brown University in 2009.

For more information contact the technical host Curt Canada, cvc@lanl.gov, 665-7453.

Hosted by the Information Science and Technology Institute (ISTI)